Social Semantic Emotion Analysis for Innovative Multilingual Big Data Analytics Markets

D4.5 Emotion Recognition from Image and Video Content, initial version

MixedEmotions

Project ref. no	H2020 644632
Project acronym	MixedEmotions
Start date of project (dur.)	01 April 2015 (24 Months)
Document due Date	30 December 2015
Responsible for deliverable	Brno University of Technology
Reply to	ihradis@fit.vutbr.cz
Document status	Final

Project reference no.	H2020 644632
Project working name	MixedEmotions
Project full name	Social Semantic Emotion Analysis for Innovative Multi Big Data Analytics Markets
Security (distribution level)	PU
Contractual delivery date	31st July 2015
Deliverable number	D4.5
Deliverable name	Emotion Recognition from Image and Video Content, initial version
Type	Report
WP / Task responsible	WP4
Contributors	BUT(Michal Hradiš), UP(Hesam Sagha)
EC Project Officer	Susan Fraser

Table of Contents

1 Executive Summary

Emotion recognition is the keystone of the MixedEmotions project. The project encompasses emotion recognition from multiple modalities (text, speech and video) and multiple languages. This document, Deliverable 4.5, describes available tools, methods and datasets for extracting high-level information from images and videos of people. This includes the analysis of facial expressions and underlying emotions, of deceit and honesty, and of personal traits such as gender, age, race, and personality. Further, in video clips where no face or human appears (such as scene of war), detection of emotions are not straightforward and face analysis is not applicable. Therefore, other approaches should be taken into account. In such videos, detection of violence can benefit emotion recognition through their correlations or by fusion of their results. For example, a video clip with high violence is expected to have low Valence and high Arousal in AV emotional space. In this deliverable we also introduce a method on violence detection from video clips.

2 Introduction

The first part of this document (Sections 3 and 4) focuses on scenarios where one person sits and either passively observes some audio/video content or participates in face-to-face or video-mediated interaction with other people. The static scenarios include consumer studies of, for example, advertisements, product designs, movie trailers, and TV shows. The dynamic scenarios include various interviews and possibly video-mediated communication and collaboration. The document presents methods and available tools which are needed to build an emotion recognition system for the intended scenarios. The methods and tools expect that at least one near-frontal view of the studied person is available, showing the face and, if possible, the full upper body.

In order to obtain meaningful high-level semantic information such as emotions, basic information about the position and orientation of body parts has to be obtained. This document discusses available tools for face detection, facial part localization, head orientation estimation, and upper body part localization (Sections 3.1, 3.2, 3.3, 3.5). These low-level methods are well studied and the available tools provide reasonable quality of results.

Personal information (e.g. gender, age, race, and possibly personality) and appearance-related information (e.g. facial hair, glasses) can be estimated from static facial images using existing pattern recognition methods trained on publicly available datasets (Section 3.6). Similarly, estimators of facial expressions (neutral, smile, frown) can be learned from existing image and video datasets (Section 3.7). However, many of these datasets are acted which slightly limits generalization of the learned models to unrestricted and spontaneous situations where facial expressions tend to be more subtle and ambiguous. All these methods perform better on well-aligned facial images. Ideally, the faces should be "frontalized" (transformed as if directly facing a camera). The text briefly presents the available tools face alignment and frontalization.

Current state-of-the-art in emotion recognition significantly lags behind facial expression recognition. One reason is that emotion recognition is a much harder problem in which the underlying emotion has to be inferred from facial expressions while taking into account contextual and temporal information. On the other hand facial expressions are directly observed. Additionally, the complexity of emotion recognition reflects in small size or low quality of available datasets. As an alternative to high-level emotion information, medium-level features can be extracted from videos. These include encoded body, head, eye, and facial (FACS encoding) movements. Such information can be later used as features for a number of recognition tasks including emotion recognition.

An interesting topic is deceit and honesty estimation from video (Section 4). Compared to emotions as they are normally understood, deceit can be objectively assessed and it can be even a-priory known when recording data – a subject can be explicitly tasked with telling a lie or truth. Automatic deceit detection from video and similar modalities has been previously studied with encouraging results showing accuracy higher than that of human judges. The results are encouraging especially considering the small sizes of existing datasets.

Automatic deceit and honesty estimation from video can be very useful for in many application targeted by MixedEmotions. Among others, it can help process customer studies which utilize interviews and video responses to predefined questions. For these reason we propose to create a crowd-sourced audio/video dataset for deceit detection.

The second part of this document (Section 5) introduced a method to detect violence in video clips. The violence detection in movies has wide applications in assessing content of videos and its suitability for younger population. However, the level of violence could be correlated with the emotions in the video clip. For example, high violence is expected to have low Valence. The proposed method in Section 5 could be useful for emotion recognition either through the developed approaches or through the fusion of the results to boost the performance of another emotion recognizer.

3 Available tools for emotion recognition from face and body gestures

3.1 Face detection

Face detection is the most basic building block for automatic human understanding. First practical and real-time face detectors appeared after 2001 starting with the frontal face detector of (Viola and Jones 2001). State-of-the-art face detectors provide high-quality real-time and multi-view detection, and are considered mature technology. The main competing methods for face detection are boosted cascades inspired by (Viola and Jones 2001), deformable part models, and emerging convolutional neural networks. These methods provide similar results and the detection quality is determined more by used training data and implementation details rather than the method family (Benenson et al. 2014; Mathias et al. 2014).

We have several real-time face detectors available, including those implemented in-house at Brno University of Technology (Herout, Hradiš, and Zemčík 2012; Juránek et al. 2015), and other state-of-the-art detectors (Mathias et al., 2014; Dollár et al., 2009).

Figure 1: Examples of facial landmark localization using a Convolutional Network regressor trained on AFLW dataset at Brno University of Technology inspired by (Sun et al., 2013).

3.2 Facial landmark localization

Facial landmarks (eyes, nose, mouth, and exact positions of their parts) are important cues for understanding of faces and can be further used for head pose estimation, frontalization, identifications and many other tasks.

Two basic families of approaches for facial landmark localization exist. One family of methods uses geometrical and appearance models of faces which are fitted to a particular image. These models can be both 2D and 3D. Examples of such methods are various versions of Active Appearance Models, and methods using local part detectors together with probabilistic shape models (Belhumeur et al. 2013)

The other family of methods is purely appearance-based without an explicit shape model. An example of such methods is the Deep Convolution Network Cascade of Sun, Wang, and Tang (2013).

At Brno University of Technology, we have replicated the work of Sun, Wang, and Tang (2013) with good results when train on the Annotated Facial Landmarks in the Wild (AFLW) dataset (Kostinger et al. 2011). Examples of facial landmark localization using this method are shown in Figure 1.

The Annotated Facial Landmarks in the Wild (AFLW) dataset (Kostinger et al. 2011) contains over 24,000 real-world images of faces in various poses gathered from Flickr. The images are annotated with up to 21 facial landmarks and with roll, pitch, and yaw describing the head pose.

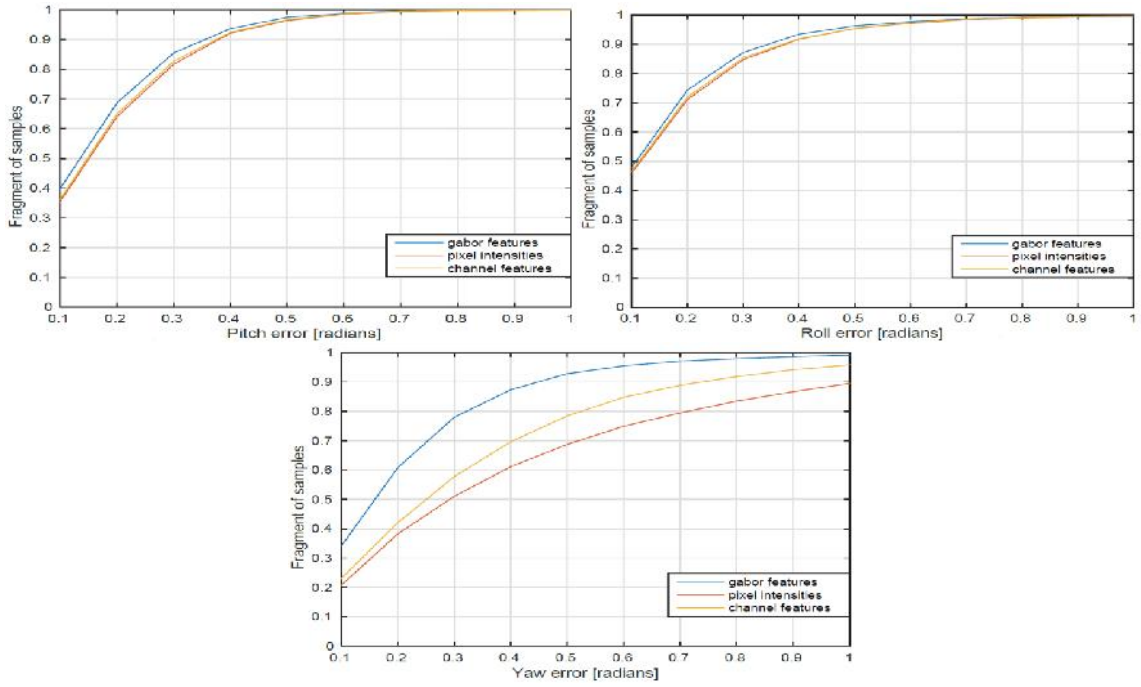

Figure 2: Cumulative error histograms of head pose estimation using random regression forests and three different features on a subset of AFLW dataset (5000 images). The y-axis represents the fraction of samples with error (yaw, pitch, and roll angles) less or equal to the value on the x-axis.

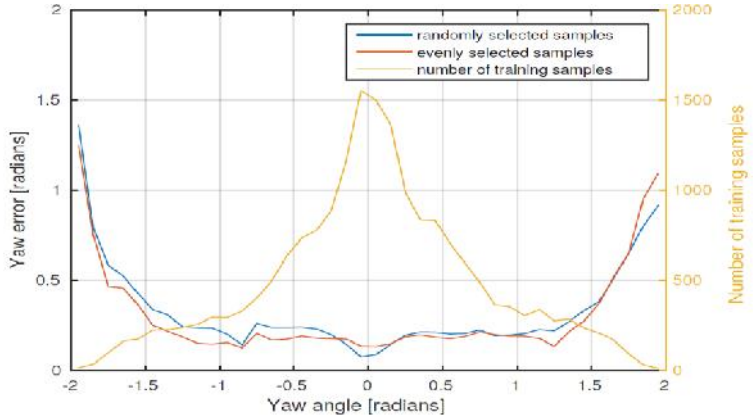

Figure 3: Error distribution over the range of yaw angle. The *yellow curve* represents the distribution of the training samples, *blue curve* represents the error when the training samples were selected randomly, and the *red curve* represents the error on samples with more evenly distributed yaw.

3.3 Head pose estimation

Head pose estimation from images is considered a mature technology at least for good quality images and frontal or profile views. It can be computed using a geometrical model form facial landmarks or it can be estimated directly from images by appearance-based classifiers. At Brno University of Technology, we developed two appearance-based methods which are available for MixedEmotion project. Both methods were trained on AFLW dataset. One of the methods uses convolutional networks and the other uses Random Regression Forests.

The Random Regression Forests head estimation was originally developed for A-PiMod (Behúň, Herout, and Pavelková 2015). It was trained on AFLW dataset (Kostinger et al. 2011). Although, the head poses in AFLW are estimated from the 21 hand-annotated facial landmarks which may introduce small errors in the ground truth, the estimated orientations are precise enough for most practical applications.

The head pose estimation accuracy was initially evaluated on a random subset of the AFLW dataset and it is shown in Figure 2. For this experiment, 50 individual trees with depth of 12 were used on three different types of features: pixel intensities, integral channel features (Dollár et al. 2009), and Gabor wavelets (Daugman 1985). Figure 2 shows errors in yaw estimation for different ground truth yaw angles. Clearly, the estimation is much more precise for near-frontal views and worse for near-profile views. One of the reasons for this behavior is that the AFLW dataset contains mostly near-frontal views (distribution of poses is shown as the yellow curve in Figure 3). When the same random forest regressor was trained on a balanced dataset, the overall error decreased.

The Random Regression Forests head estimation was verified in A-PiMod project on a custom dataset with ground truth pose information obtained using OptiTrack motion tracking system[1]. Examples from the dataset are shown in Figure 4. The yaw angle results for one video from the dataset, together with ground truth data from the OptiTrack sensor, are shown in Figure 5, and overall results from multiple video sequences are summarized in Table 1.

[1] https://www.naturalpoint.com/optitrack/

Figure 4: Collected simulator dataset for head pose estimation evaluation. In the second row, the head position is drawn into the images together with lines representing the ground truth for yaw and pitch angles (green lines) and the angles estimated by the Random forests (blue, red and yellow lines, for near-frontal, near-profile left and near-profile right detectors, respectively). The TrackClipPRO with three LEDs, which was used for tracking to get the ground truth data, is visible on the left side of the subjects' heads.

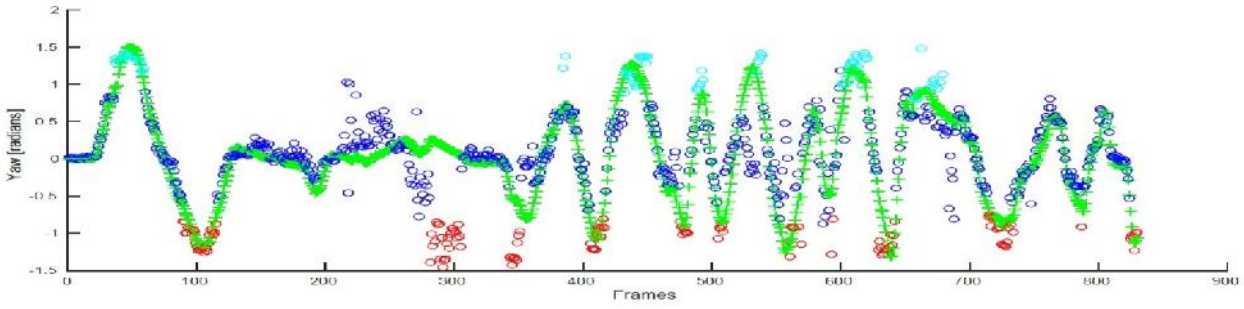

Figure 5: Quality of head pose estimation on a single video sequence. The ground truth (green '+') and estimated values of yaw angle (blue, cyan and red circles) for one video.

	Yaw	Pitch	Roll
Overall pose error (2RF)	0.1418	0.2316	0.0664
Near-frontal pose error (2RF)	0.0994	0.2309	0.0627
Near-profile pose error (2RF)	0.3286	0.2374	0.0825
Overall pose error (1RF)	0.1479	0.2313	0.0667

Table 1: Median of head pose estimation error in radians for video with OptiTrack ground truth data.

3.4 Face Alignment and Frontalization

Face frontalization is an important part of many methods which extract information from faces. The goal of face frontalization is to create an image with a face in a canonical frontal pose which would be as similar as possible to an observed image. Frontalization greatly reduces appearance variability due to pose changes which is irrelevant to tasks such as identification and estimation of age, gender, and expression. For example the state-of-the-art person recognizer DeepFace (Taigman et al. 2014) heavily relies on frontalization. The presented results suggest that approximately half of improvement of DeepFace comes from high-quality frontalization and half from the large Convolutional Neural Network which processes the frontalized images.

A simpler version of frontalization is face alignment which applies 2D geometric transformations to facial images such that they would overlap as well as possible. Facial alignment can be performed based on localized facial landmarks or it can be performed jointly directly on the images in unsupervised way (Huang et al., 2012).

While frontalization can be considered mostly experimental, tools for facial alignment are mature, reliable, and easy to use. Currently, we use translation, scale and rotation alignment based on estimated positions of eyes and mouth.

3.5 Body pose estimation

Body pose estimation in the context of this document is a 2D localization of large human body parts in images and in video. Body pose estimation from images and video is a hot research topic and the state-of-the-art methods, in general, provide reasonably precise localization of body parts in good quality images and common poses. Most pose estimation methods model appearance of body parts and distribution of their relative positions in 2D. These methods include various part-based model with Pictorial Structure (Dantone et al. 2014). An interesting variant of such approach is the Pose Machine by (Ramakrishna et al. 2014)

which replaces the graphical model of relative part positions with regression trees and applies them recursively on their own body part estimations.

At Brno University of Technology in A-PiMod project (Behúň et al. 2015), we have developed a variant of the Pose Machine which is tuned to upper body pose estimation. It is able to localize 10 important upper body joints: *head*, *shoulder center*, *right/left shoulder*, *right/left elbow*, *right/left wrist*, *right/left hand*. In our experiments, the Pose Machine achieved superior results compared to other state-of-the-art solutions of human pose estimation such DeepPose (Szegedy, Toshev, and Erhan 2013) a Pictorial Structure (Dantone et al. 2014).

The principle of Pose Machine is shown in Figure 6. Pose machine is a hierarchical method consisting of multiple stages. Each stage is modelled by a multiclass random forest with produces position probability maps for each body part. The position probability maps produced by one stage of Pose Machine is used as in input for the next stage as spatial context features which allow the random forest to learn relationships among body parts and thus improve pose estimation accuracy for each body part. In fact, infinitely deep Pose Machine is equivalent to inference a fully-connected graphical model of the body part positions; however, the random forests with context features are able to capture much stronger dependencies among the parts and their appearance which the graphical models used in pictorial structures cannot.

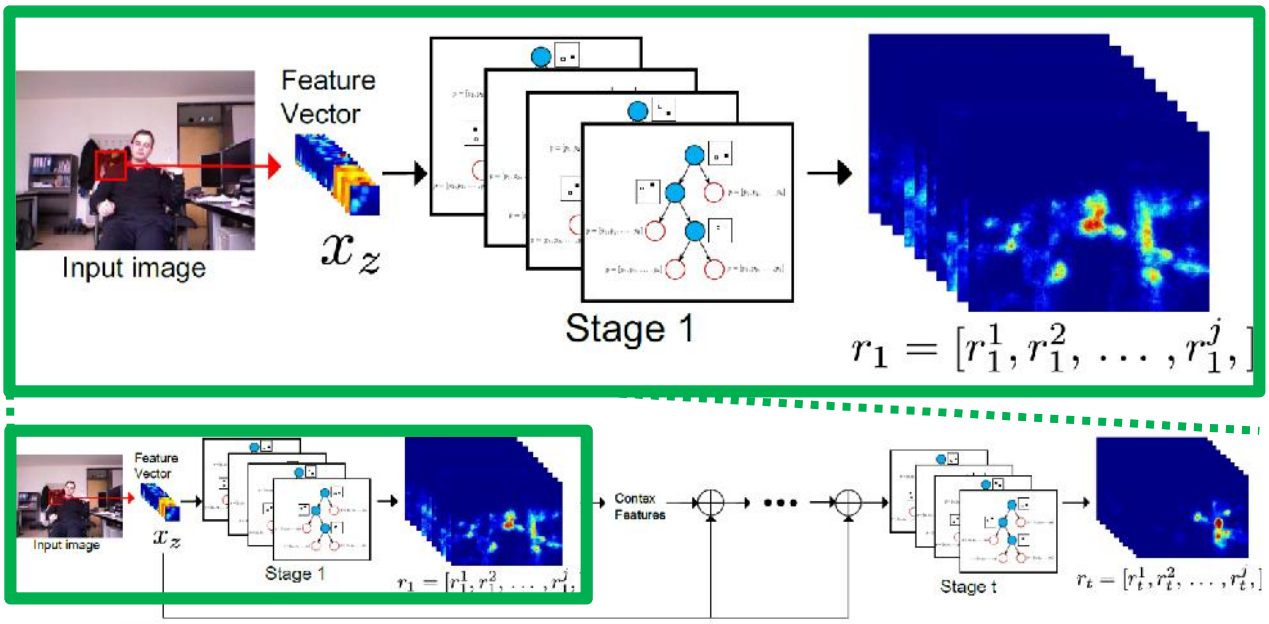

Figure 6: Human pose estimation by hierarchical inference machine named Pose Machine. Input of stage t includes features computed from input image and context features computed from output of stage $t-1$ for each body part. Each stage is modeled by a random forest.

The human pose estimation based on Pose Machines was trained on a dataset which contains videos of 24 seated people in an office environment (Behúň, Herout, and Páldy 2014) (see Figure 7). The dataset contains 6,213 pose frames at 640x480 resolution with manually verified positions of 10 upper body joints. The same set of features (except for the output of a skin detector) was used as in the work by (Dantone et al. 2014) (resulting in 16 feature channels). Each stage contained 15 trees and the Pose Machine reached the best achievable performance when using 3 stages. The results are shown in Figure 8.

For videos, the per-frame joint position estimations can be smoothed using for example a particle filter. The A-PiMod dataset is fairly small and, as a consequence, the Pose Machines struggle to learn general body part appearance. If needed, the accuracy of this method could be significantly improved by increasing the size of the training dataset.

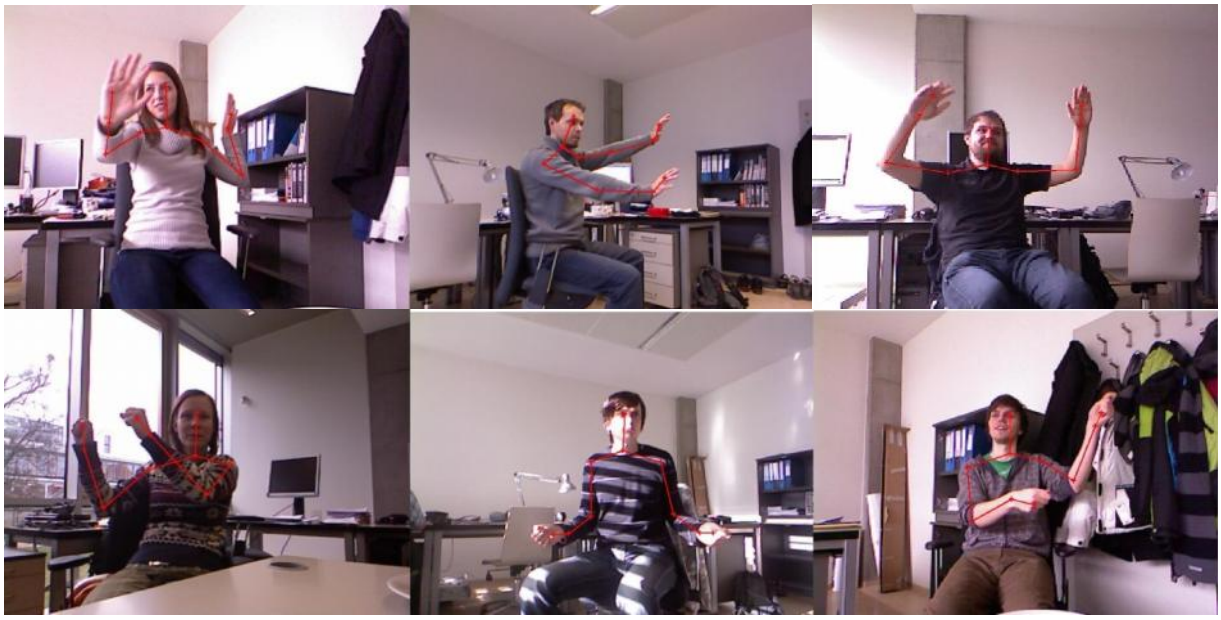

Figure 7: Examples from the A-PiMod pose estimation dataset with annotations.

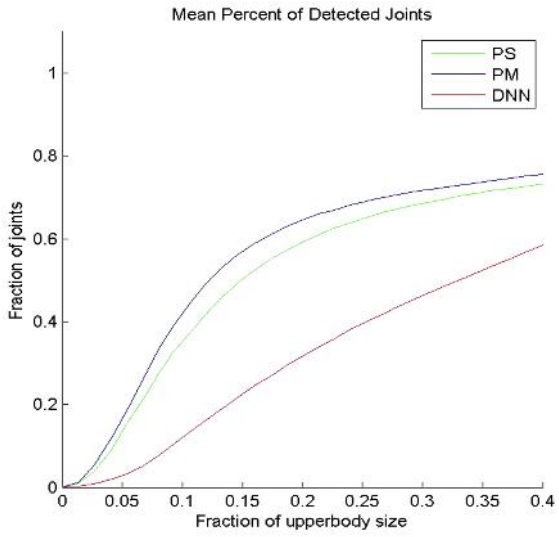

Figure 8: Pose estimation results on the A-PiMod dataset. Three methods are compared – Convolutional Neural Network regression (DNN), Pictorial Structure (PS), and Pose Machines (PM). The graph shows what fraction of body parts is correctly located within increasing tolerance. The error tolerance is expressed as fraction of upper body size (distance between opposite shoulder and hip).

3.6 Personal and appearance traits

The default approach to estimation of personal and appearance traits from facial images is to classify aligned or frontalized facial images with state-of-the-art classification methods. Currently, the state-of-the-art classifiers for images are based on Convolutional Neural Networks in task ranging from semantic segmentation, object detection, and general photo classification to person identification and attribute modeling. We follow this trend and use convolutional neural network for many image pattern recognition tasks. Using a unified approach with CNN allows us to use the same tools in various applications which reduces development time and cost while providing excellent performance and quality.

Possibilities of estimation of different personal and appearance traits directly depends on the available datasets and on their annotation. Table 2 summarizes some of the existing dataset with trait annotations. The datasets provide enough data for estimation of age, gender race, facial hair, hair style, eye wear, and even subjective attributes such as attractiveness.

When working with videos instead of images, per-frame estimations can be easily integrated into a single result for the full video resulting in increased accuracy.

Name	#Images	#People	Environment	Traits	Notes
10k US Adult Faces Database	2,222		In the wild	Age, facial hair, gender, race, attractiveness, emotion, eye direction, face direction, …	
Unfiltered faces for gender and age classification	26,580	2,284	In the wild	Age, gender, anonymized identity	
Large-scale CelebFaces Attributes	202,599	10,177	In the wild	bald, hair color, eyeglasses, makeup, facial hair, skin color, smiling, hair style, hat, young, …	Celebrity images from the web
FaceScrub	107,818	530	In the wild	Gender	Celebrity images from the web
Face Tracer	15,000	--	In the wild	5000 face labels - gender, race, age, hair color, eye wear, mustache, smiling.	
Pub Fig	58,797	200	In the wild	Neutral / non-neutral expression	

Table 2. Existing datasets with annotations of personal and appearance traits.

3.7 Facial expressions and emotions

Automatic recognition of emotions from video has wide applications for example in human-computer interaction (Brown 2014), and recommendation systems (Berkovsky 2015). However the research in automatic emotion recognition is hampered by the available datasets which are usually focused on a specific scenario or situation and existing methods trained on such datasets can't be expected to generalize well to other applications (e.g. RECOLA database (Sonderegger, Sauer, and Lalanne 2013). Further, the task of emotion recognition is inherently hard as reliable ground truth is very hard to create due to the hidden nature and complexity of emotions. The available approaches and their performance on the RECOLA database is summarized in document D4.7 Emotion Recognition from Multilingual Audio Content, initial version.

The RECOLA database contains spontaneous and naturalistic interactions collected during the resolution of a collaborative task that was performed in dyads and remotely through video conference. Multimodal signals, i.e., audio, video, electro-cardiogram (ECG) and electro-dermal activity (EDA), were synchronously recorded from 27 French-speaking subjects. The recordings were annotated with time-continuous ratings (40 *ms* binned frames) of emotional arousal and valence by six gender-balanced French-speaking assistants for the first five minutes of all recordings. In the initial five minutes, participants discussed more about their strategy – hence showing emotions – at the beginning of their interaction.

Facial expressions are one of the observable expressions of emotions (most of the time). The expressions are directly observable and quantifiable (Ekman and Friesen 1978), and as such are much easier to automatically detect compared to emotions. Basic facial expressions can be detected similarly to the appearance traits with reasonable accuracy using existing publicly available datasets. Such expression detectors are constrained by the existing datasets which contain mostly frontal faces and basic facial expressions (neutral, smile, sad, angry, surprise).

Alternatively FACS codes (Ekman and Friesen 1978) or similar could be extracted from images and used further as features.

4 Deceit dataset

Automatic deceit recognition would be useful in many scenarios which MixedEmotions targets. For example honesty of online video review could be automatically assessed and dishonest survey respondents could be detected. Such automatic approaches would enable larger-scale operations, and it could even improve reliability of obtained information as existing research suggests that people are notoriously bad at jugging deceitful behavior, and automatic systems often surpass even trained humans in this task (Meservy et al. 2005; Michael et al. 2010; Perez-Rosas et al. 2014; Yap et al. 2011; Yu et al. 2015).

Compared to emotions, deceit can be assessed much more easily – it can be even know a-priory as participants can be explicitly asked to respond honestly or dishonestly in predefined way.

The goal would be to collect a large-scale real-life dataset suitable for deceit detection with scenarios relevant to possible real-life applications in video-mediated communication or video recorded responses to surveys and questionnaires. The desired size of the dataset is about 500 interviews each 20 minutes long.

4.1 Recording setup

We propose to record participants in their own environment using their own equipment in order to obtain as natural recordings as possible. Such requirements limit the available recording equipment to basic video and audio. We intend to instruct the participants to use a high-resolution web camera with built-in microphone positioned right in front of them in such a way that it would record full upper body as shown in Figure 9. Pre-recorded video questions and statements would be played on a computer screen in a web browser and the video and audio would be recorded in a browser as well.

A smaller number of sessions would be recorded in a lab with extended sensors including multiple cameras, depth camera, and high-quality microphones.

Figure 9: Trial recordings for the proposed deceit dataset.

4.2 Participants

Participants will be hired using Mechanical Turk and by personal contacts. The language of the interviews will be English, but the participants would not be required to be native speakers. Their English proficiency should be sufficient to allow them to study or work in an English a speaking country. Participants will be incentivized to participate through a small payment per recording and a prize draw organized at the end.

4.3 Preliminary scenarios

Existing deceit datasets contain small number of recurring scenarios. The most common is mock theft scenario.

Few exemplar scenarios which we consider are:

- Alibi – This scenario should simulate interrogation by a law enforcement officer. The participant has to recall events from previous day and change his narrative in specific way.
- Work interview – Starting with a short questionnaire designed to asses participant's real working experience, he or she is instructed to improve his existing work experience or to add completely new work experience.

- Personal relations – the participant is instructed to describe his friend or colleague which he has good relations with, and a person with whom he has bad relations. Further, the participant has to describe his friend in negative way and the other person as positive as possible.
- After a short questionnaire designed to assess participants' ethical or political views, the participant is asked to defend or explain his position on one of the topics which he or she has strong opinion about. Consequently, he or she is asked to defend the opposite view.

4.3.1 Preliminary script for alibi scenario

We have recorded pilot sessions for the Alibi scenario in a lab. This recording was fully automatic with pre-recorded instructions and questions. Only very brief instructions how and where to sit were given to the participants orally.[2] The script which was used in the pilot sessions was very similar to the following one:

1. (Video) Hello and welcome. The recordings of this session will be used for a research of verbal and non-verbal communication. Please, try to make your statements as plausible as possible and present them in a convincing way. We will evaluate the perceived honesty of your statements and the three best participants will receive 100 euro as a reward.
2. (Video) Imagine that you are sitting in a cold and unfriendly interrogation room at a police station. You have been sitting here for the last twenty minutes waiting for a policeman investigating a deadly accident that happened yesterday. Use the next minute to recall what you did yesterday with as many details as possible. Remember that you will be required to answer policeman's questions afterward.
3. (1 minute pause)
4. (Video) Please, select a period approximately two hours long from yesterday in which you did not spend at home. All following question will target this specific time period. Describe the selected time period in the next two minutes. Focus on your activities and your immediate surroundings.
5. (1 minute answer RECORDING)
6. (Video) Thanks, that is enough. In the next 1 minute, elaborate
    a. who you interacted with during the time period and who could remember you.
    b. how did you get to the place your previous description started with.
7. (1 minute answer RECORDING)
8. (Video) Thanks, that is enough. Now think how to incorporate a 30 minute trip to a supermarket into your previous story and you bought more than three items.
9. (1 minute pause)

[2] Videos from the three pilot sessions are available at https://drive.google.com/open?id=0B0j386T9FUjwbHVHX1lacThTYUk

10. (Video) Describe the period from the start, this time including the made up trip to the supermarket. Keep in mind that you are trying to convince a policemen that you went really visited the supermarket. Focus on your activities and situation in your surroundings. You have the next two minutes to record your statement.
11. (1 minute answer RECORDING)
12. (Video) Thanks, that is enough. In the next 1 minute, please elaborate
    a) who you interacted with during your trip to the supermarket and who could remember you.
    b) what way you took to the supermarket and what kind of transport you used.

    (1 minute answer RECORDING)

    (Video) Thanks. How much did you pay with a card? (20s for answer)
    (Video) Thanks. What color of hair did the cashier have? (20s for answer)
    (Video) Thanks. How long did you wait in the queue? (20s for answer)
    (Video) Thanks. Did you weight the fruit you bought yourself? (20s for answer)
    (Video) Thanks. Which item took you the longest to find? (20s for answer)
    (Video) Thanks. Did you take a trolley or a shopping basket? (20s for answer)
    (Video) Thanks. Did you pay with a card? (20s for answer)
    (Video) Thanks. How much did you pay? (20s for answer)
    (Video) Thanks. What exactly did you buy? (20s for answer)
    (Video) Thank you very much for your time and effort. (20s for answer)

5 Violence detection in movie

In videos which facial or body gesture estimations cannot be applied (i.e., videos with no person), other approaches should be considered to recognize emotions. As one primary solution, in the following, we introduce our method on violence detections in movie clips. We expect that violence has a great correlation with some emotions (such as anger), therefore, it may help emotion recognition.

Our method is the Imperial College London, Technische Universität München and University of Passau (ICL+TUM+PASSAU) team approach to the MediaEval's "Affective Impact of Movies" challenge, which consists in the automatic detection of affective (arousal and valence) and violent content in movie excerpts. In addition to the baseline features, we computed spectral and energy related acoustic features, and the probability of various objects being present in the video. Random Forests, AdaBoost and Support Vector Machines were used as classification methods. Best results show that the dataset is highly challenging for both affect and violence detection tasks, mainly because of issues in inter-rater agreement and data scarcity.

5.1 Introduction

The MediaEval 2015 Challenge "Affective Impact of Movies" comprises two subtasks using the LIRIS-ACCEDE database (Baveye et al. 2015). Subtask 1 targets the automatic categorization of videos in terms of their affective impact. The goal is to identify the arousal (calm-neutral-excited) and valence (negative-neutral-positive) levels of each video. The goal of Subtask 2 is to identify those videos that contain violent scenes. The full description of the tasks can be found in (Sjöberg et al. 2015).

5.2 Methodology

5.2.1 Subtask 1: affect classification

5.2.1.1 Feature sets

In our work we have used both the baseline features provided by the organisers (Baveye et al. 2015), as well as our own sets of audio-visual features as described below. The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) was used to extract acoustic features with the openSMILE toolkit (Eyben, Weninger, Groß, et al. 2013); this feature set was designed as a standard acoustic parameter set for automatic speech emotion recognition (Eyben et al. 2015; Ringeval et al. 2014; Ringeval, Valstar, et al. 2015) and has also been successfully used for other paralinguistic tasks (Ringeval, Marchi, et al. 2015). The eGeMAPS comprises a total of 18 Low-Level Descriptors (LLDs), including frequency, energy/amplitude, and spectral related features. Various functionals were then applied to the LLDs over the whole instance, giving rise to a total of 88 features.

The emotional impact of videos can be heavily influenced by the kind of objects present in a given scene (Hanjalic and Xu 2005; Hu et al. 2011; Lopatovskaa and Arapakis 2011). We thus computed a probability of 1000 different objects to be present in a frame using a pretrained 16-layer convolutional neural network (CNN) on the ILSVRC2013 dataset (Deng et al. 2009; Simonyan and Zisserman 2014). Let $\chi \in \mathbb{R}^{N \times p}$ represent a video of the database with *N* frames and *p* pixels per frame, and *f(.)* the trained convolutional neural net with *softmax* activation functions in the output layer. The probability $\Pr(y = c|x_i; \theta)$ for each of the 1000 classes being present inside the *i*-th frame of a video x_i is obtained by forwarding the *p* pixels

value through the network. By averaging the activations over all the *N* frames of a video sequence we obtained the probability distribution of the 1000 ILSVRC2013 classes that might be present in the video.

5.2.1.2 Classifiers

For modelling the data we concentrated on two out-of-the-box ensemble techniques: Random Forests and AdaBoost. We used these two techniques as they are less susceptible to the overfitting problem than other learning algorithms due to the combination of weak learners, they are trivial to optimise as they have only one hyper-parameter, and they usually provide close or on par results with the state-of-the-art for a multitude of tasks (G., Gall, and L. 2011; Guo et al. 2011; Lee et al. 2013; Stumpf and Kerle 2011). The hyper-parameters for each classifier were determined using a 5-fold cross-validation scheme on the development set. During development the best performance was achieved with 10 trees with Random Forests and 20 trees with AdaBoost.

5.2.1.3 Runs

We submitted a total of five runs. Run 1 consisted of predictions using the baseline features and the AdaBoost model. The predictions in runs 2 and 5 were obtained using the baseline plus our audio-visual feature sets and the Random Forest and AdaBoost classifiers, respectively. By looking at the distribution of labels in the development set, we observed that the most common combinations of labels are: 1) neutral valence (V^n) and negative arousal (A^-) (24%), and 2) positive valence (V^+) and negative arousal (A^-) (20%). Runs 3 and 4 are thus based on the hypothesis that the label distribution of the test set will be similarly unbalanced. In run 3 every clip was predicted to be V n, A+ and in Run 4 every one was V^+, A^-. These submissions act as a sanity check of our own models, but also other competitors' submissions for this competition.

5.2.2 Subtask 2: violence detection

5.2.2.1 Feature sets

According to previous work (Eyben, Weninger, Lehment, et al. 2013; Ionescu et al. 2013), we only considered spectral and energy based features as acoustic descriptors. Indeed, violent segments do not

necessarily contain speech; voice specific features, such as voice quality and pitch related descriptors, might thus not be a reliable source of information for violence. We extracted 22 acoustic low-level descriptors (LLDs): loudness, alpha ratio, Hammarberg's index, energy slope and proportion in the bands [0-500]Hz and [500-1500]Hz, and 14 MFCCs, using the openSMILE toolkit (Eyben, Weninger, Groß, et al. 2013). All LLDs, with the exception of loudness and the measures of energy proportion, were computed separately for voiced and unvoiced segments. As the frames of the movie that contain violent scenes are unknown, we computed 5 functionals (max, min, range, arithmetic mean and standard-deviation) to summarise the LLDs over the movie excerpt, which provided a total of 300 features. For the video modality, we used the same additional features defined in Subtask 1. We also used the metadata information of the video genre as an additional feature, due to dependencies between movie genre and violent content.

5.2.2.2 Classifier

Since the dataset is strongly imbalanced – only 272 excerpts out of 6,144 are labelled as violent – we up-sampled the violent instances to achieve a balanced distribution. All features were furthermore standardised with a z-score. As classifier, we used the *libsvm* implementation of Support Vector Machines (SVMs) (Chang and Lin 2011) and optimised the complexity parameter, and the γ coefficient of the radial basis kernel in a 5-folds cross-validation framework on the development set. Because the official scoring script requires the computation of *a posteriori* probabilities, which is more time consuming than the straightforward classification task, we optimised the Unweighted Average Recall (UAR) to find the best hyper-parameters (Schuller et al. 2014, 2015), and then re-trained the SVMs with the probability estimates.

5.2.2.3 Runs

We first performed experiments with the full baseline feature set and found that the addition of the movie genre as feature improved the Mean Average Precision (MAP) from 19.5 to 20.3, despite degrading the UAR from 72.3 to 72.0. Adding our own audio-visual features provided a jump in the performance with the MAP reaching 33.6 and UAR 77.6. Because some movie excerpts contain partly relevant acoustic information, we empirically defined a threshold on loudness based on the histogram, to exclude frames before computing the functionals. This procedure has improved the MAP to 35.9 but downgraded the UAR to 76.9. A fine tuning of the complexity parameter and coefficient yielded the best performance in terms of

UAR with a value of 78.0, but slightly deteriorated the MAP to 35.7. We submitted a total of five runs. Run 1 – baseline features; Run 2 – all features mentioned above (except movie genre) with loudness threshold (0.038); Run 3 – same as Run 2 plus the inclusion of movie genre; Run 4 – as Run 3 but with a fine tuning of the hyper-parameters; Run 5 – similar to Run 3 but with a higher threshold for loudness (0.078).

5.3 Results

Our official results on the test set for both subtasks are shown in Table 3.

Subtask 1: Our results for the affective task indicate that we did not do much better than was expected by chance for arousal classification, and did slightly better than chance for valence in run 5; we thus refrain from further interpretation of results. This can be explained by the low quality of the provided annotations for the dataset. The initial annotations had a low inter-rater agreement (Baveye et al. 2015), and there were multiple processing stages afterwards (Baveye et al. 2014; Sjöberg et al. 2015) with high levels of uncertainty and unclear validity.

Subtask 2: Results show that there is an important over-fitting in our models as the performance is divided by a factor of 2 between development and test partitions. This is, however, not really surprising since only 272 instances labelled as violent were available as training data. Moreover, the labelling task being performed not at the frame level but rather at the excerpt level does not allow to model precisely the information that is judged as violent, making the task highly challenging. We can nevertheless observe that the proposed audio-visual feature set brings a large improvement over the baseline feature set – the MAP is improved by a factor superior to 2, and that the inclusion of the movie genre as additional feature also allows a small improvement in the performance.

Table 3. The submission results for the arousal, valence, and violence classification tasks on the test partition. AC stands for accuracy and MAP for the mean average precision.

	Subtask1		Subtask 2
Run	Arousal (AC)	Valence (AC)	Violence (MAP)
1	55.72	39.99	4.9
2	54.71	41.00	13.3
3	55.55	37.87	13.5
4	55.55	29.02	14.9
5	54.46	41.48	13.9

5.4 Discussion

We have presented our approach to violence detection in movies. It consists in the automatic detection of affective and violent content in movie excerpts. Our results for the affective task have shown that we did not do much better than a classifier that is based on chance, although we use features and classifiers that are known to work well in the literature for arousal and valence prediction (Baveye et al. 2015; Forbes-Riley and Litman 2004). We consider that this might be owed to a potentially noisiness of the annotations provided. As for the violence prediction subtask, the results show that we over-fit a lot on the development set, which is not very striking given the small amount of instances of the minority class. The analysis of violent content at the excerpt level is also highly challenging, because only few frames might contain violence, and such brief information is almost totally lost in the computation of functionals at the full excerpt level.

6 Conclusions

MixedEmotions focuses on extraction of emotion from audio, video and text. This deliverable summarizes the necessary available tools and methods for emotion recognition from video content. The document presented available tools which can be used as building blocks for emotion recognition from faces: face detection, head pose estimation, facial landmark localization, and alignment. These basic tools enable

further personal and appearance trait estimation including facial expressions and emotions. Pose Machines are able to estimate upper body pose and movement which can serve as additional cue for emotion recognition. In addition, detection of violence in movie clips is our primary step toward emotion recognition in videos where face and body gestures are not tractable.

References

Baveye, Yoann, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. 2014. “From Crowdsourced Rankings to Affective Ratings.” Pp. 1–6 in *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW 2014)*.

Baveye, Yoann, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. 2015. “LIRIS-ACCEDE: A Video Database for Affective Content Analysis.” *IEEE Transactions on Affective Computing* 6(1):43–55.

Behúň, Kamil, Adam Herout, and Alexander Páldy. 2014. “Kinect-Supported Dataset Creation for Human Pose Estimation.” Pp. 88–95 in *Proceedings of Spring Conference on Computer Graphics*. Comenius University in Bratislava. Retrieved (http://www.fit.vutbr.cz/research/view_pub.php?id=10674).

Behúň, Kamil, Adam Herout, and Alena Pavelková. 2015. “Hand Gestures in Aeronautics Cockpit as a Cue for Crew State and Workload Inference.” Pp. 1–6 in *Proceedings of ITSC 2015*. The Universidad de Las Palmas de Gran Canaria. Retrieved (http://www.fit.vutbr.cz/research/view_pub.php?id=10906).

Belhumeur, Peter N., David W. Jacobs, David J. Kriegman, and Neeraj Kumar. 2013. “Localizing Parts of Faces Using a Consensus of Exemplars.” Pp. 2930–40 in *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 35.

Benenson, R., M. Omran, J. Hosang, and and B. Schiele. 2014. “Ten Years of Pedestrian Detection, What Have We Learned?” in *ECCV, CVRSUAD workshop*.

Berkovsky, Shlomo. 2015. “Emotion-Based Movie Recommendations: How Far Can We Take This?” P. 1 in *Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems 2015*.

Brown, Stephen. 2014. “Meet Pepper, The Emotion Reading Robot.” *TECHNOLOGY*.

Dantone, M., J. Gall, C. Leistner, and Luc Van Gool. 2014. “Body Parts Dependent Joint Regressors for Human Pose Estimation in Still Images.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* (99):1. Retrieved (http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6802375\npapers3://publication/doi/10.1109/TPAMI.2014.2318702).

Daugman, John G. 1985. “Uncertainty Relation for Resolution in Space, Spatial Frequency, and Orientation

Optimized by Two-Dimensional Visual Cortical Filters." *Journal of the Optical Society of America. A, Optics and image science* 2(7):1160–69.

Deng, Jia et al. 2009. "Imagenet: A Large-Scale Hierarchical Image Database." Pp. 248–55 in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*.

Dollár, Piotr, Zhuowen Tu, Pietro Perona, and Serge Belongie. 2009. "Integral Channel Features." in *BMVC 2009*. Retrieved March 11, 2012 (http://dblp.uni-trier.de/db/conf/bmvc/bmvc2009.html#DollarTPB09).

Ekman, Paul, and Wallance V. Friesen. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*.

Eyben, F. et al. 2015. "The {Geneva} Minimalistic Acoustic Parameter Set ({GeMAPS}) for Voice Research and Affective Computing." *IEEE Transactions on Affective Computing* In press.

Eyben, Florian, Felix Weninger, Florian Groß, et al. 2013. "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor." Pp. 835–38 in *Proceedings of the 21st ACM International Conference on Multimedia (MM 2013)*. Barcelona, Spain.

Eyben, Florian, Felix Weninger, Nicolas Lehment, Björn Schuller, and Gerhard Rigoll. 2013. "Affective Video Retrieval: Violence Detection in Hollywood Movies by Large-Scale Segmental Feature Extraction." *PLOS one* 8(12):1–12.

Forbes-Riley, Katherine, and Diane J. Litman. 2004. "Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources." Pp. 201–8 in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (HLT-NAACL)*.

G., Fanelli, J. Gall, and Van Gool L. 2011. "Real Time Head Pose Estimation with Random Regression Forests." Pp. 617–24 in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. Providence (RI), USA.

Guo, Li, Nesrine Chehata, Clément Mallet, and Samia Boukir. 2011. "Relevance of Airborne Lidar and Multispectral Image Data for Urban Scene Classification Using Random Forests." *ISPRS Journal of Photogrammetry and Remote Sensing* 66(1):56–66.

Hanjalic, Alan, and Li-Qun Xu. 2005. "Affective Video Content Representation and Modeling." *IEEE Transactions on Multimedia* 7(1):143–54.

Herout, Adam, Michal Hradiš, and Pavel Zemčík. 2012. "EnMS: Early Non-Maxima Suppression." *Pattern Analysis and Applications* 2012(2):121–32. Retrieved (http://www.fit.vutbr.cz/research/view_pub.php?id=9506).

Hu, Weiming, Nianhua Xie, Li Li, Xianglin Zeng, and Maybank S. 2011. "A Survey on Visual Content-Based Video Indexing and Retrieval." *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 41(6):797–819.

Chang, Chih-Chung, and Chih-Jen Lin. 2011. "LIBSVM: A Library for Support Vector Machines." *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3):27. Retrieved February 18, 2012 (http://dl.acm.org/citation.cfm?id=1961189.1961199).

Ionescu, Bogdan, Jan Schlüter, Ionut Mironica, and Markus Schedl. 2013. "A Naive Mid-Level Concept-Based Fusion Approach to Violence Detection in Hollywood Movies." Pp. 215–22 in *Proceedings of the 3rd ACM International Conference on Multimedia Retrieval (ICMR 2013)*. Dallas (TX), USA.

Juránek, Roman, Adam Herout, Markéta Dubská, and Pavel Zemčík. 2015. "Real-Time Pose Estimation Piggybacked on Object Detection." in *ICCV*.

Kostinger, M., P. Wohlhart, P. M. Roth, and H. Bischof. 2011. "Annotated Facial Landmarks in the Wild: A Large-Scale, Real-World Database for Facial Landmark Localization." Pp. 2144–51 in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*.

Lee, Jung-Jin, Pyoung-Hean Lee, Seong-Whan Lee, Alan Yuille, and Christof Koch. 2013. "AdaBoost for Text Detection in Natural Scene." Pp. 429–34 in *Proceedings of the IEEE 12th International Conference on Document Analysis and Recognition (ICDAR 2013)*. Beijing, China.

Lopatovskaa, Irene, and Ioannis Arapakis. 2011. "Theories, Methods and Current Research on Emotions in Library and Information Science, Information Retrieval and Human--Computer Interaction." *Information Processing & Management* 47(4):575–92.

Mathias, M., R. Benenson, M. Pedersoli, and L. Van Gool. 2014. "Face Detection without Bells and Whistles." in *ECCV*.

Meservy, T. O., M. L. Jensen, J. Kruse, J. K. Burgoon, and J. F. Nunamaker. 2005. "Automatic Extraction of Deceptive Behavioral Cues from Video." *Intelligence and Security Informatics, Proceedings* 3495:198–208.

Michael, Nicholas, Mark Dilsizian, Dimitris Metaxas, and Judee K. Burgoon. 2010. "Motion Profiles for

Deception Detection Using Visual Cues." *Proceedings 11th European Conference on Computer Vision* 462–75.

Perez-Rosas, Veronica, Rada Mihalcea, Alexis Narvaez, and Mihai Burzo. 2014. "A Multimodal Dataset for Deception Detection." *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* 3118–22. Retrieved (http://www.lrec-conf.org/proceedings/lrec2014/summaries/869.html).

Ramakrishna, Varun, Daniel Munoz, Martial Hebert, J. Andrew Bagnell, and Yaser Sheikh. 2014. "Pose Machines : Articulated Pose Estimation via Inference Machines." Pp. 1–15 in *ECCV 2014*.

Ringeval, Fabien, Erik Marchi, et al. 2015. "Face Reading from Speech -- Predicting Facial Action Units from Audio Cues." Pp. 1977–81 in *Proceedings INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association (ISCA)*. Dresden, Germany: ISCA.

Ringeval, Fabien, Shahin Amiriparian, Florian Eyben, Klaus Scherer, and Björn Schuller. 2014. "Emotion Recognition in the Wild: Incorporating Voice and Lip Activity in Multimodal Decision-Level Fusion." Pp. 473–80 in *Proceedings of the ICMI 2014 EmotiW -- Emotion Recognition In The Wild Challenge and Workshop (EmotiW 2014), Satellite of the 16th ACM International Conference on Multimodal Interaction (ICMI 2014)*. Istanbul, Turkey: ACM.

Ringeval, Fabien, Michel Valstar, Erik Marchi, Denis Lalanne, and Roddy Cowie. 2015. "The AV + EC 2015 Multimodal Affect Recognition Challenge : Bridging Across Audio , Video , and Physiological Data Categories and Subject Descriptors." *Proc. ACM Multimedia Workshops* (CCC):2–5.

Schuller, Björn et al. 2014. "The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load." Pp. 427–31 in *Proceedings INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association*. Singapore, Singapore: ISCA.

Schuller, Björn et al. 2015. "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson's & Eating Condition." in *Proceedings of INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association (ISCA)*. Dresden, Germany.

Simonyan, Karen, and Andrew Zisserman. 2014. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *arXiv preprint arXiv:1409.1556*.

Sjöberg, M. et al. 2015. "The MediaEval 2015 Affective Impact of Movies Task." Pp. 14–16 in *Proceedings of the MediaEval 2015 Workshop*. Wurzen, Germany.

Sonderegger, F. Ringevaland A., J. Sauer, and D. Lalanne. 2013. “Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions.” in *Proc. Face and Gestures 2013, Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE)*.

Stumpf, André’, and Norman Kerle. 2011. “Object-Oriented Mapping of Landslides Using Random Forests.” *Remote Sensing of Environment* 115(10):2564–77.

Sun, Yi, Xiaogang Wang, and Xiaoou Tang. 2013. “Deep Convolutional Network Cascade for Facial Point Detection.” Pp. 3476–83 in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*.

Szegedy, Christian, Alexander Toshev, and Dumitru Erhan. 2013. “Deep Neural Networks for Object Detection.” Pp. 2553–61 in *Advances in Neural Information Processing Systems 26*, edited by C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger. Retrieved (http://papers.nips.cc/paper/5207-deep-neural-networks-for-object-detection.pdf).

Taigman, Yaniv, Ming Yang, Marc’Aurelio Ranzato, and Lior Wolf. 2014. “DeepFace: Closing the Gap to Human-Level Performance in Face Verification.” Pp. 1701–8 in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. Retrieved April 16, 2015 (http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6909616).

Viola, Paul, and Michael Jones. 2001. “Rapid Object Detection Using a Boosted Cascade of Simple Features.” *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on* 1:511.

Yap, Moi Hoon, Bashar Rajoub, Hassan Ugail, and Reyer Zwiggelaar. 2011. “Visual Cues of Facial Behaviour in Deception Detection.” *ICCAIE 2011 - 2011 IEEE Conference on Computer Applications and Industrial Electronics* (Iccaie):294–99.

Yu, Xiang et al. 2015. “Is Interactional Dissynchrony a Clue to Deception? Insights from Automated Analysis of Nonverbal Visual Cues.” *IEEE Transactions on Cybernetics* 45(3):506–20.